View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by Springer - Publisher Connector

Shan *et al. BMC Medical Research Methodology* (2017) 17:22
DOI 10.1186/s12874-017-0297-5

BMC Medical Research Methodology

RESEARCH ARTICLE Open Access

Efficient confidence limits for adaptive one-arm two-stage clinical trials with binary endpoints

Guogen Shan[1], Hua Zhang[2] and Tao Jiang[3*]

Abstract

Background: Recently, several adaptive one-arm two-stage designs have been developed by fully using the information from previous stages to reduce the expected sample size in clinical trials with binary endpoints as primary outcome. It is important to compute exact confidence limits for these studies.

Methods: In this article, we propose three new one-sided limits by ordering the sample space based on *p*-value, average response rate at each stage, and asymptotic lower limit, as compared to another three existing sample size ordering approaches based on average response rate. Among the three proposed approaches, the one based on the average response rate at each stage is not exact, and the remaining two approaches are exact with the coverage probability guaranteed.

Results: We compare these exact intervals by using the two commonly used criteria: simple average length and expected length. The existing three approaches based on average response rate have similar performance, and they have shorter expected lengths than the two proposed exact approaches although the gain is small, while this trend is reversed under the simple average criterion.

Conclusions: We would recommend the two exact proposed approaches based on *p*-value and asymptotic lower limit under the simple average length criterion, and the approach based on average response rate under the expected length criterion.

Keywords: Adaptive design, Clopper-Pearson approach, Exact one-sided interval, Response rate, Two-stage design

Background

To assess the activity of a new treatment in a cancer clinical trial, Simon's two-stage designs [1] are traditionally used among the multi-stage designs. Simon's two-stage designs can be improved by allowing the second stage sample size to depend on the number of responses observed from the first stage, which is an adaptive two-stage design [2–6]. Recently, Shan et al. [6] developed an adaptive two-stage design that meets the non-increasing relationship between the second stage sample size and the number of responses from the first stage. This is an adaptive one-arm two-stage design that allows the second stage sample size to change with the number of responses observed from the first stage, and the second stage sample size is a non-increasing function of the first stage responses. This sample size monotonic constraint is considered as an intuitive property in adaptive two-stage designs: fewer participants are needed in the second stage when more responses are observed from the previous stage. It should be noted that these adaptive designs [6] often allow early stopping in the first stage due to futility or efficacy, while Simon's design only allows stopping for futility in the first stage. In these studies, the hypothesis for testing the activity of the new treatment is often one-sided.

Upon completion of an adaptive clinical trial, it is important to provide statistical inference based on the number of responses and the number of participants in each stage. Recently, Zhao et al. [7] proposed a likelihood based approach to construct confidence interval for a study

*Correspondence: jtao@263.net
[3]Department of Statistics, Zhejiang Gongshang University, Hangzhou, 310018 Zhejiang, China
Full list of author information is available at the end of the article

BioMed Central

© The Author(s). 2017 **Open Access** This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The Creative Commons Public Domain Dedication waiver (http://creativecommons.org/publicdomain/zero/1.0/) applies to the data made available in this article, unless otherwise stated.

that is designed by Simon's two-stage method but with unplanned second stage sample size [8, 9]. The likelihood approach was shown to be associated with good performance with regard to coverage probability and coverage bias, but this interval is asymptotic. To guarantee the coverage probability, Shan [10] proposed several new exact confidence intervals based on exact binomial distribution calculation. These intervals are developed for traditional Simon's two-stage designs whose second stage sample size is considered as fixed regardless the number of responses observed from the first stage as long as it is over the threshold to move to the second stage.

In this article, we consider statistical inference for adaptive two-stage designs whose second stage sample sizes depend on the first stage responses. Adaptive designs are generally flexible and effective as compared to the traditional Simon's design, however they are often computationally due to many design parameters in the design. With the new adaptive designs being proposed, it is important to develop statistical inference for these designs. To preserve the nominal coverage probability, we propose developing exact one-sided confidence intervals for the response rate in an adaptive two-stage design setting. Confidence intervals are computed by using exact binomial distributions instead of asymptotic distributions. The hypothesis for these trials is often one-sided and the null hypothesis is rejected when a high response rate is observed. For this reason, we focus our approach on exact one-sided lower intervals. The interval is computed by using the approach developed by Clopper and Pearson [11], who proposed the commonly used exact one-sided intervals for a binomial proportion. This approach has to be used in the conjunction with a method to order the sample space.

Multiple approaches have been proposed to order the sample space for multi-stage designs [9, 12–18]. Four orderings were discussed by Jennison and Turnbull [15] after a group sequential design: stage-wise ordering, maximum likelihood estimate (MLE) ordering, likelihood ratio (LR) ordering, and score test (Score) ordering. Lower and upper confidence limits are used in the first ordering. When the outcome is binary, the MLE ordering is equivalent to the ordering by average response rate, which is the number of responses divided by the total sample size in the study. The last two orderings depend on average response rate and the number of sample size from that stage.

In addition to the three aforementioned sample space orderings for a study with binary endpoints, we propose three new methods to order the sample space. The first method is the one based on the response rates from the first and second stages. We consider this is an intuitive method for ordering the sample space based on the information from both stages. However, we find that not all sample points can be ordered by using this method. This leads to the situation in which the nominal coverage probability is not guaranteed. Although the lower limits from the first method are not exact, they can be used as a measurement to order the sample space again. In the second method, each sample point has a unique order number based on the asymptotic lower limit from the first method. The third method uses the p-value of each sample point to create a new ordering of the sample space. We find that the ordering based on the p-value has a very interesting relationship with that from the first method.

In "Methods" section, we first introduce the basic settings for adaptive two-stage designs, then propose three new methods to order the sample space. When a study is stopped in the first stage due to either futility or efficacy, their ordering positions are the same in each method. For this reason, we focus on the ordering for sample points when a trial goes to the second stage. We then investigate the coverage probability for each interval. In "Results" section, we compare the performance of the two proposed exact approaches and the three existing approaches with regards to simple average length (AL) and expected length (EL). These approaches are compared by using the completed sample space. In addition to that comparison, we also introduce a new subsample space including the sample points whose second stage response rate is within the confidence interval of the first stage response rate. In other words, if a study's first stage response rate is very different from its second stage response rate, the study population could be changed (e.g., disease status, gender ratio), and other approaches should be explored for such cases. For this reason, we also utilize this new sample space in the performance comparison among the exact intervals. Finally, we conclude our research with a discussion in "Discussion" section.

Methods

To test the activity of a treatment in a two-stage design compared to the historical response rate π_0, the hypotheses are often presented as

$$H_0 : \pi \leq \pi_0,$$

against the alternative

$$H_a : \pi \geq \pi_1,$$

where π_1 is the estimated response rate of the new treatment. The null hypothesis is rejected for a high response rate.

Simon's two-stage design is traditionally used in clinical trials to assess the activity of a new cancer treatment. In Simon's design, the second stage sample size $n_2(X_1)$ is a constant when a trial goes to the second stage, where X_1 is the number of responses from the first stage. From an adaptive perspective, a clinical trial could be much more

flexible and effective when the second stage sample size is allowed to change based on information gathered so far. It is reasonable to assume that $n_2(X_1)$ has a non-increasing relationship with the number of responses observed from the first stage: $n_2(X_1) \geq n_2(X_1^{'})$ when $X_1 < X_1^{'}$. Shan et al. [6] developed an adaptive optimal two-stage design with the non-increasing sample size relationship respected. This design is referred to be as the AdaptiveS design. The design is given as

$$(n_1, n_2(X_1), r(X_1)),$$

where the number of possible responses out of n_1 participants in the first stage is $X_1 = 0, 1, 2, \ldots, n_1$, and $n_2(X_1)$ and $r(X_1)$ are the second stage sample size and the critical value for the study given X_1 responses from the first stage, respectively. In this article, a study can be stopped in the first stage due to futility when $X_1 \leq r_1(f)$ or efficacy when $X_1 \geq r_1(e)$. When the number of responses from the first stage is between $r_1(f)$ and $r_1(e)$, $r_1(f) < X_1 < r_1(e)$, the trial proceeds to the second stage with an additional $n_2(X_1)$ participants and the final decision is made by comparing the total number of responses $(X_1 + X_2)$ and $r(X_1)$, where X_2 is the number of responses out of $n_2(X_1)$ participants. The new treatment is considered effective enough to proceed to the next phase when $X_1 + X_2 \geq r(X_1)$. Otherwise, the new treatment is not promising for further investigation.

Upon completion of an adaptive clinical trial, confidence interval for the response rate should be computed and reported. The hypothesis for testing the activity of the new treatment is often one-sided, and the confidence interval and the hypothesis testing should be consistent with each other. For this reason, we focus our interest on one-sided lower intervals as the null hypothesis is rejected when a high response rate is observed. When the significance level is α, a $1 - \alpha$ one-sided interval, $(L, 1]$, should be computed for statistical inference, where L is the $1 - \alpha$ lower limit.

The method by Clopper and Pearson [11] (CP) was used to construct exact one-sided intervals for a binomial proportion. It is exact because the coverage probability is guaranteed to be at least $1 - \alpha$ and the coverage probability is calculated by using the binomial probabilities, not asymptotic distributions. We extend this approach for the response rate in adaptive two-stage design settings. This approach has to be used with a method to order the sample space, which is often referred to as stochastic ordering. The complete sample space can be divided into three complementary sub-spaces,

$$\Omega = \{G_1, G_2, G_3\},$$

where $G_1 = \{X_1 : 0, 1, 2, \ldots, r_1(f)\}$, $G_3 = \{X_1 : r_1(e), r_1(e)+1, \ldots, n_1\}$, and $G_2 = \{(X_1, X_2) : r_1(f) < X_1 < r_1(e), X_2 \leq n_2(X_1)\}$. Sets G_1 and G_3 contain the sample points where a trial is stopped in the first stage due to futility and efficacy, respectively. Set G_2 represents the sample points that a trial goes to the second stage. It should be noted that set G_3 could be empty in some cases when the optimal adaptive two-stage stops in the first stage only due to futility, which often occurs in cases with a large π_0 and a large difference between π_0 and π_1 as seen in Shan et al. [6]. The lower limits for sample points in set G_1 are the smallest, followed by sample points in set G_2 and set G_3. Within sets G_1 and G_3, the lower limits for the sample points are ordered by the number of their responses, and they are the same as the CP lower limits for a binomial proportion. For sample points in set G_2, the second stage sample size changes as the number of responses from the first stage. For this reason, the second stage sample size should be considered in the sample ordering. We propose ordering the sample points in set G_2 by response rates (RR) in the first stage and the two stages combined,

$$L\left(\frac{X_1}{n_1}, \frac{X_1 + X_2}{n_1 + n_2(X_1)}\right) \leq L\left(\frac{X_1^{'}}{n_1}, \frac{X_1^{'} + X_2^{'}}{n_1 + n_2\left(X_1^{'}\right)}\right)$$
$$\text{if } \frac{X_1}{n_1} \leq \frac{X_1^{'}}{n_1} \text{ and } \frac{X_1 + X_2}{n_1 + n_2(X_1)} \leq \frac{X_1^{'} + X_2^{'}}{n_1 + n_2\left(X_1^{'}\right)}.$$

This approach is referred to as the RR approach. This ordering is motivated by the p-value calculation for a two-stage study. The rejection region includes the extreme outcomes whose first stage and second stage responses are at least as large as the observed data [8–10, 15].

Another stochastic ordering is based on the p-value of each sample point. Similar to the RR approach, the p-value for sample points in set G_1 is the largest, followed by sample points in set G_2 and G_3. A sample point with a large p-value indicates week evidence against the null hypothesis. In other words, it should have a large lower limit. For a sample point (X_1, X_2) in set G_2, its associated p-value is calculated as

$$P(X_1, X_2) = \sum_{(X_1^{'}, X_2^{'}) \in \Theta(X_1, X_2)} b\left(X_1^{'}, n_1, \pi_0\right) b\left(X_2^{'}, n_2\left(X_1^{'}\right), \pi_0\right),$$

where $b(\ldots)$ is the probability density function of a binomial distribution, and $\Theta(X_1, X_2)$ is the tail area

$$\Theta(X_1, X_2) = \left\{ G_3 \text{ and } \left(X_1^{'}, X_2^{'}\right) : \frac{X_1}{n_1} \leq \frac{X_1^{'}}{n_1}, \frac{X_1 + X_2}{n_1 + n_2(X_1)} \leq \frac{X_1^{'} + X_2^{'}}{n_1 + n_2\left(X_1^{'}\right)} \right\}.$$

The response rates are used to define the tail area in the p-value calculation. Since the p-value is used to order the sample space in this approach, we name this approach as the PV approach. Although the p-value calculation may not guarantee the type I error rate, it is still a valid measurement to order the sample space. In this approach, we

sort the sample points by the p-value from smallest to largest. In the PV approach, every sample point has its order number based on its p-value, from 1 to the size of the sample space. In the RR approach, two sample points from set G_2 can be ordered only if one sample point belongs to another's tail area.

Once a stochastic ordering of the sample space is defined, we use the CP method to compute the exact one-sided lower limit as the collection of π

$$\left\{\pi : P\left(\Omega_{\varphi}(X_1, X_2)|\pi\right) > \alpha\right\}, \tag{1}$$

where φ is an approach used to order the sample space (e.g., PV, RR), $\Omega_{PV}(X_1, X_2) = \left\{\left(X_1^{'}, X_2^{'}\right) : P\left(X_1^{'}, X_2^{'}\right) \leq P(X_1, X_2)\right\}$, and $\Omega_{RR}(X_1, X_2) = \left\{\left(X_1^{'}, X_2^{'}\right) : \left(X_1^{'}, X_2^{'}\right) \in \Theta(X_1, X_2)\right\}$. Since the null hypothesis is rejected for a large response rate, we focus on the one-sided lower limit. The proposed approach to compute exact one-sided lower limits can be readily applied to calculate exact upper limits.

Theorem 1 *For any given response rate* π, $P\left(\Omega_{PV}(X_1, X_2)|\pi\right) \geq P\left(\Omega_{RR}(X_1, X_2)|\pi\right)$ *is always true.*

Proof For sample points from set G_1 and set G_3, it is easy to show that $P\left(\Omega_{PV}(X_1, X_2)|\pi\right)$ is always equal to $P\left(\Omega_{RR}(X_1, X_2)|\pi\right)$. For a given sample point $\left(X_1^{'}, X_2^{'}\right) \in \Omega_{RR}(X_1, X_2)$ where sample points (X_1, X_2) and $\left(X_1^{'}, X_2^{'}\right)$ are from set G_2, the relationship between their tail areas is

$$\Theta(X_1, X_2) \supseteq \Theta\left(X_1^{'}, X_2^{'}\right).$$

Thus, the p-value of (X_1, X_2) is not less than that of $\left(X_1^{'}, X_2^{'}\right)$: $P(X_1, X_2) \geq P\left(X_1^{'}, X_2^{'}\right)$. It follows that the sample point $\left(X_1^{'}, X_2^{'}\right)$ belongs to $\Omega_{PV}(X_1, X_2)$. Therefore, $P\left(\Omega_{PV}(X_1, X_2)|\pi\right)$ is always greater than or equal to $P\left(\Omega_{RR}(X_1, X_2)|\pi\right)$ □

From Theorem 1 and the construction of the exact one-sided interval in Eq. (1), we see that the exact one-sided lower limit based on the RR approach is always greater than or equal to that based on the PV approach.

Coverage probability is defined as

$$P\Big(\pi \in (L(X_1, X_2), 1]\Big) = P\Big((X_1, X_2) : L(X_1, X_2) < \pi|\pi\Big). \tag{2}$$

A confidence interval is called exact if $P(\pi \in (L(X_1, X_2), 1]) \geq 1 - \alpha$ is satisfied for any $\pi \in [0, 1]$. We present the coverage probability plots for the adaptive design with design parameters $(\pi_0, \pi_1, \alpha, \beta) = (20\%, 40\%, 0.05, 0.2)$ [6] in Fig. 1. It can be seen that the PV approach is exact with the coverage probabilities being at least 95%. However, the RR approach is not exact, as the coverage probability could be as low as 90.5% at the nomial level of 95% in this configuration. One reason for the non-exactness of the RR approach is that the sample points can not be completely ordered. To overcome this issuee, we use the non-exact lower limits from the RR approach to order the sample space again. A new stochastic ordering is created by using the calculated limits. This new ordering can be viewed as a two-step ordering because the ordering is generated after the non-exact limit calculation. This approach is referred to as the RR-A approach.

The following three existing approaches to order the sample space have been discussed in the literature [15]. Sample space can be ordered by the average response rate from the study,

$$\frac{X_1 + X_2}{n_1 + n_2(X_1)}.$$

This approach is referred to as the RR-B approach. It should be noted that this sample size ordering is equivalent to the ordering by using the MLE in a one-sample problem [15]. Another existing sample size ordering is based on the LR method, which is given as

$$\frac{X_1 + X_2}{n_1 + n_2(X_1)}\sqrt{n_2(X_1)},$$

named as the RR-LR approach. Similar to the RR-LR approach, Rosner and Tsiatis [14] discussed another ordering based on the score test:

$$\frac{X_1 + X_2}{n_1 + n_2(X_1)}n_2(X_1).$$

This approach is referred to be as the RR-Score approach. In general, we would expect a significant number of ties when only the response rate is used to order the sample space in traditional two-stage designs. However, the number of ties could be reduced in adaptive design settings as the second stage sample size $n_2(X_1)$ is a non-increasing function of X_1, not a constant.

Results

We first compare the performance of the existing three approaches (RR-B, RR-LR, RR-Score) for sample size orderings. The sample space is first ordered by these approaches, then the CP method is used to obtain exact one-sided limits. Figure 2 shows the interval length comparison among the three approach for the adaptive two-stage design with design parameters $(\pi_0, \pi_1, \alpha, \beta) = (30\%, 50\%, 0.05, 0.1)$, where the interval length is defined as $[1 - L(X_1, X_2)]$ for each sample point (X_1, X_2) from the

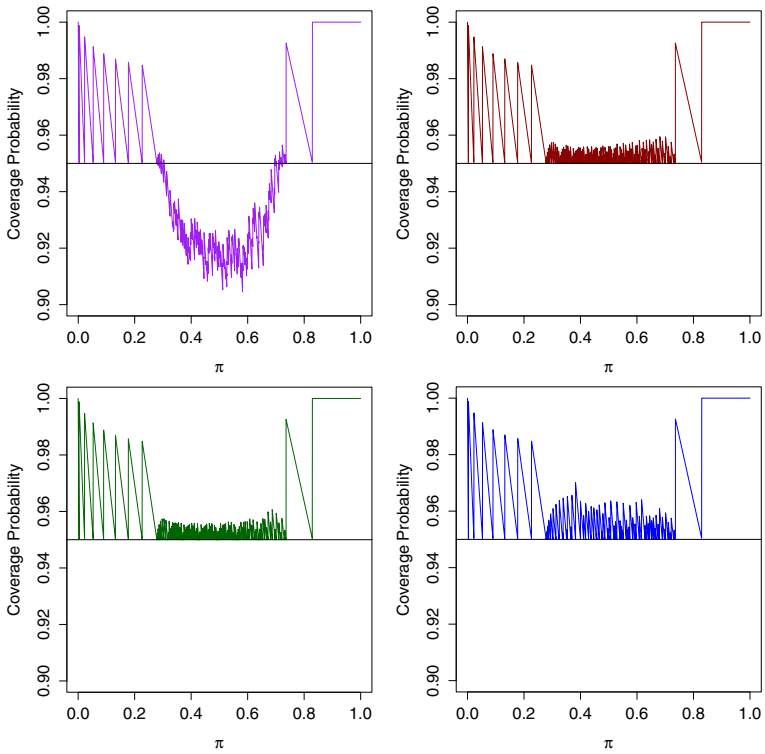

Fig. 1 Coverage probability for 95% one-sided lower intervals of the four approaches for the AdaptiveS two-stage design with design parameters $(\pi_0, \pi_1, \alpha, \beta) = (20\%, 40\%, 0.05, 0.2)$. RR approach: *top left*; PV approach: *top right*; RR-A approach: *bottom left*; RR-B approach: *bottom right*

sample space. The overall average is almost identical for these three approaches. It can be seen from the figure that their interval lengths are very close to each other. Given the simplicity of the RR-B approach, we would recommend that to be used in practice as compared to the RR-LR approach and the RR-Score approach.

We present the coverage probability for the adaptive design with design parameters $(\pi_0, \pi_1, \alpha, \beta) = (20\%, 40\%, 0.05, 0.2)$ in Fig. 1. It can be seen that the RR-A approach, the RR-B approach, and the PV approach guarantee the coverage probability while the RR approach does not. For this reason, the RR approach is not going to be included in the following performance comparision. We have illustrated that these three approaches are exact with coverage probability guaranteed. Two criteria, simple average length and expected length, are used to compare the performance of these exact limits. As aforementioned, set G_1 and set G_3 represent a study being stopped in the first

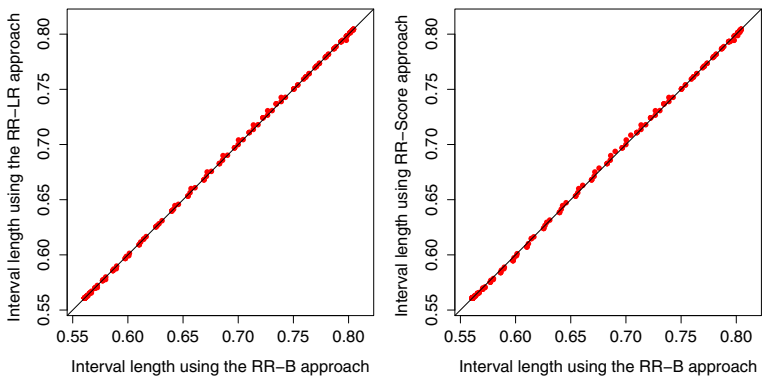

Fig. 2 Confidence interval length comparison for each sample point from the adaptive two-stage design with design parameters $(\pi_0, \pi_1, \alpha, \beta) = (30\%, 50\%, 0.05, 0.1)$ when the AdaptiveS design is used. The RR-B approach, the RR-LR approach, and the RR-Score approach are compared

stage due to futility or efficacy, respectively. Their lower limits are the same for the three approaches, and actually, they are the exact intervals based on the CP method for a binomial proportion. For this reason, we exclude these sample points in the performance comparison.

The sample space for set G_2 is given as $G_2 = \left\{(X_1, X_2) : r_1(f) < X_1 < r_1(e), 0 \le X_2 \le n_2(X_1)\right\}$, and the size of this set is

$$M = \sum_{X_1=r_1(f)+1}^{r_1(e)-1} (n_2(X_1) + 1).$$

The simple average length is defined as

$$AL = \frac{\sum_{(X_1,X_2)\in G_2} \left[1 - L(X_1, X_2)\right]}{M}.$$

We use 16 different adaptive optimal two-stage designs with π_0 from 5 to 70%, $\pi_1 = \pi_0 + 20\%$, 80% or 90% power, and $\alpha = 0.05$ from Shan et al. [6], in the perfomrance comparison among the three approaches. For each adaptive two-stage design, we first calcualte the exact one-sided lower limits for each sample point in the sample space G_2, then the AL value is computed. The AL comparison among the three approaches for the 16 configurations is presented on the left side of Fig. 3. An approach with a shorter average length is preferable. Thus, the approach on the y-axis performs better than the approach on the x-axis when a point is below the diagnoal line. Each point in the plot represents the AL values for an adaptive design. It can be seen that the RR-A approach and the PV approach often have shorter average lengths than the RR-B approach, and the RR-A approach and the PV approach are generally comparable. We also investigate another 16 configurations with $\pi_1 = \pi_0 + 15\%$ as in Shan et al. [6], and we observe similar conclusions. We compute the overall average of AL values for these 32 configurations. The RR-A approach has the shortest overall average length (0.6057), followed by the PV approach (0.6069), and the RR-B approach (0.6273).

It should be noted that some sample points in set G_2 may not be practically possible, for example, a sample point with a relatively very low or high estimated second stage response rate as compared to the first stage response rate. From a practical perspective, it would be reasonable to expect a study with similar response rates from each stage. For this reason, we create a new subset of G_2 to compare the perform again, and the new subset includes the sample points whose second stage response rates are within the 95% confidence interval of their first stage response rates. The confidence interval of the first stage response rate is computed by the function *exactci* from R package *PropCIs* [19]. The CP method is used in the function *exactci* to obtain exact two-sided intervals for a binomial proportion. The new subset of G_2 is defined as $G_2(CI)$

$$G_2(CI) = \left\{(X_1, X_2) : r_1(f) < X_1 < r_1(e), 0 \le X_2 \le n_2 (X_1), X_2 \in CI(X_1, n_1, 95\%)\right\},$$

where $CI(X_1, n_1, 95\%)$ is the 95% exact two-sided interval for the response rate X_1/n_1. The simple average length for sample points in set $G_2(CI)$ is computed as

$$AL = \frac{\sum_{(X_1,X_2)\in G_2(CI)} \left[1 - L(X_1, X_2)\right]}{M(CI)},$$

where $M(CI)$ is the size of sample space $G_2(CI)$. The three approaches are also compared for the 16 adaptive designs when the sample space is $G_2(CI)$, in Fig. 3. It should be noted that $G_2(CI)$ is the same for the three approaches as it only depends on the first stage response rate, not on the approach. After removing the non-practical sample points, the three approaches have very close average lengths. By using the total of 32 configurations as aforementioned, the overall average length when using the subsample space $G_2(CI)$ is 0.6018, 0.6028, and 0.6038 for the RR-A approach, the PV approach, and the RR-B approach, respectively. It suggests that the RR-A approach has the best performance when $G_2(CI)$ is considered. We include the sample points whose second stage response rate are within the 95% confidence interval of their first stage response rates. If the current considered confidence level of 95% is decreased to 80% or increased to 97.5%, we observe similar results. When it is increased to 99% with more sample points in the sample space, the results are similar to these observed from the case when G_2 is the sample space. When the confidence level is increased to 100%, the sample space $G_2(CI)$ is the same as G_2.

In addition to the average length comparison among the three approaches, we also present the interval length comparison between all the sample points, $[1 - L(X_1, X_2)]$, in Fig. 4 for the adaptive two-stage design with design parameters $(\pi_0, \pi_1, \alpha, \beta) = (30\%, 50\%, 0.05, 0.1)$. When all sample points from G_2 are considered, the RR-B approach has longer lengths than the other two approaches in general. However, the three approaches are similar to each other for sample points from the subset space $G_2(CI)$. The advantage of the RR-A approach and the PV approach over the RR-B approach is mostly due to the sample points that are not practically possible. The RR-A approach and the PV approach are comparable under either sample space. Similar results are observed from other configurations, and other existing adaptive two-stage designs that do not meet the monotonic sample size property [20].

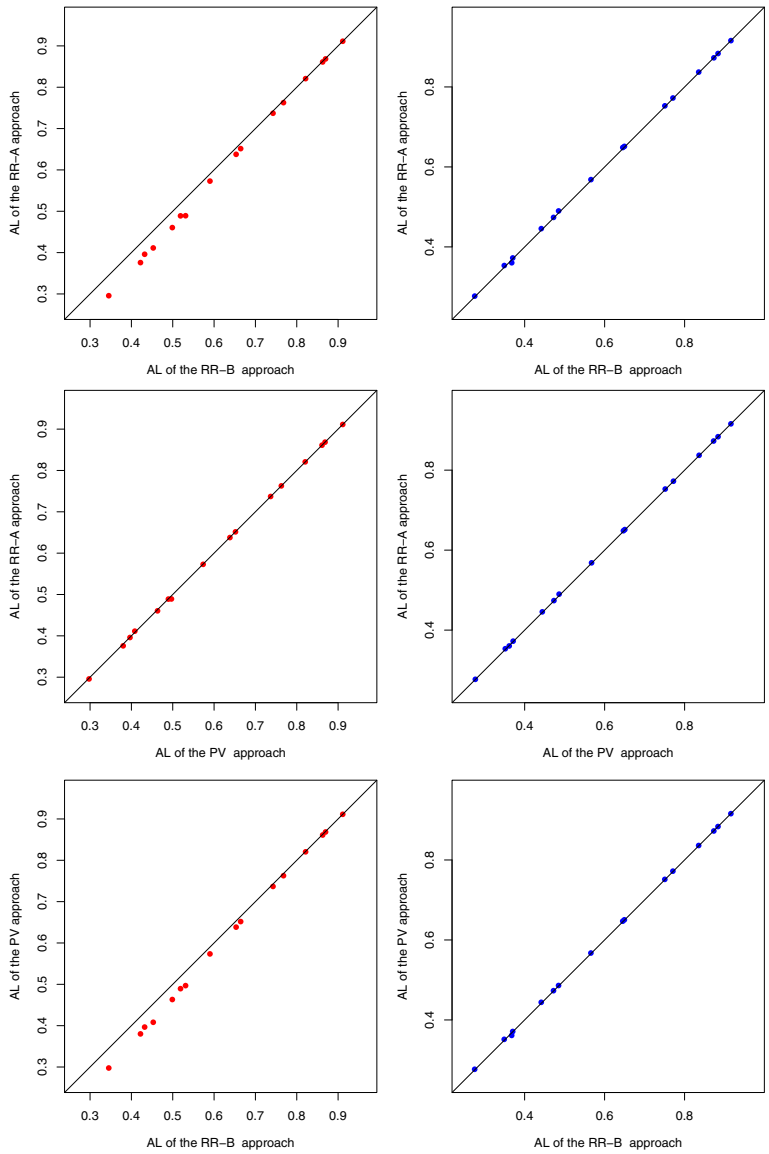

Fig. 3 Average length comparison among the three exact approaches (the RR-A approach, the RR-B approach, and the PV approach), for the 16 different configurations with $\pi_1 = \pi_0 + 20\%$ when the AdaptiveS design is used. The sample space G_2 (*left side*) and the subsample space $G_2(CI)$ (*right side*) are used in computing the 95% one-sided exact lower limits. An approach with a lower average length is preferable

Under the simple average length criterion, each sample point's interval length in the sample space is weighted equally. It is also important to compare the performance of exact limits by using the expected length criterion. Under this criterion, each sample point's interval length is weighted by its associated probability at the response rate π. Thus, the EL is a function of π. Specifically, the EL is defined as

$$EL(\pi) = \sum_{(X_1,X_2)\in G_2} [1-L(X_1,X_2)]\, b(X_1,n_1,\pi)b(X_2,n_2(X_1),\pi),$$

where $b(X_1,n_1,\pi)b(X_2,n_2(X_1),\pi)$ is the probability of the observed data (X_1,X_2) with n_1 and $n_2(X_1)$ as the first stage and the second stage sample sizes. For a given π, we first compute the $EL_{RR-A}(\pi)$, $EL_{RR-B}(\pi)$, and $EL_{PV}(\pi)$. Their differences are generally small, especially in cases where π is near the boundary. For this reason, we use their ratios to compare them. Figure 5 displays the EL ratio plots for the adaptive two-stage design with design parameters $(\pi_0,\pi_1,\alpha,\beta) = (20\%, 40\%, 0.05, 0.1)$. In general, these three approaches are comparable with regards to the EL criterion, with the EL ratio between 99.5% and 100.6%. When set G_2 is the sample space, the RR-B approach has a shorter expected length than the other two approaches when $\pi > 20\%$. When the sample space $G_2(CI)$ is used, the

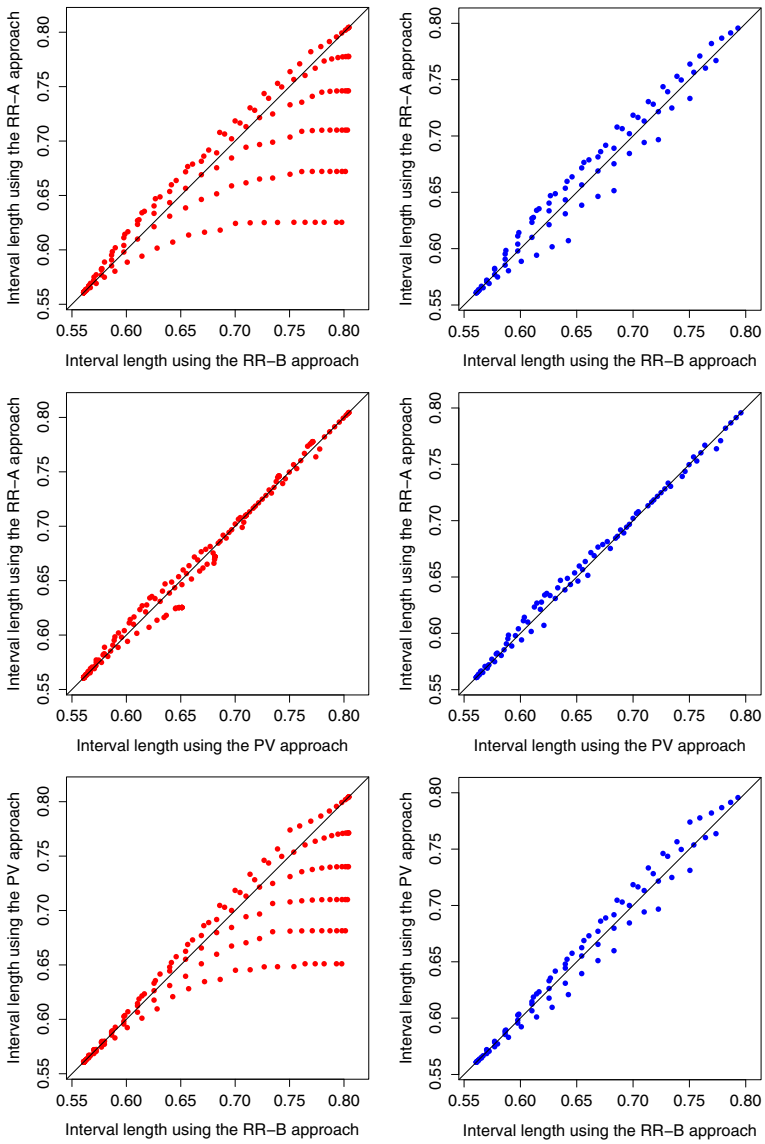

Fig. 4 Confidence interval length comparison for each sample point from the adaptive two-stage design with design parameters $(\pi_0, \pi_1, \alpha, \beta) = (30\%, 50\%, 0.05, 0.1)$ when the AdaptiveS design is used. The RR-A approach, the RR-B approach, and the PV approach are compared on the left side when the sample space G_2 (*left side*) is used, and on the right side when the subsample space $G_2(CI)$ (*right side*) is used

RR-B appraoch generally performs better than the RR-A approach and the PV approach, although the gain is small. The RR-A approach has a shorter expected length than the PV approach when π is near $\pi_0 = 20\%$. Similar results are observed when other adaptive designs are used.

Example

Since the considered adaptive two-stage designs are relatively new, we do not expect to find a real data set to be used. For this reason, we assume that a study is designed by using the AdaptiveS design with design parameters $(\pi_0, \pi_1, \alpha, \beta) = (30\%, 50\%, 0.05, 0.1)$. In the first stage, $X_1 = 11$ responses are observed among the $n_1 = 22$ patients. Then, the required number of patients is $n_2(X_1) = 35$ for the second stage. At the end of the study, we assume that $X_2 = 13$ responses are observed from the second stage. Thus, the total number of responses is $11 + 13 = 24$, and the average response rate is estimated as $24/(22 + 35) = 42.1\%$. The 95% lower limit for the response rate is calcualted as 0.320 for the PV approach, 0.325 for the RR-A approach, 0.317 for the RR-B approach, 0.317 for the RR-LR approach, 0.317 for the RR-Score approach, and 0.344 for the RR approach. It can be seen that the RR-B approach, the RR-LR approach, and the

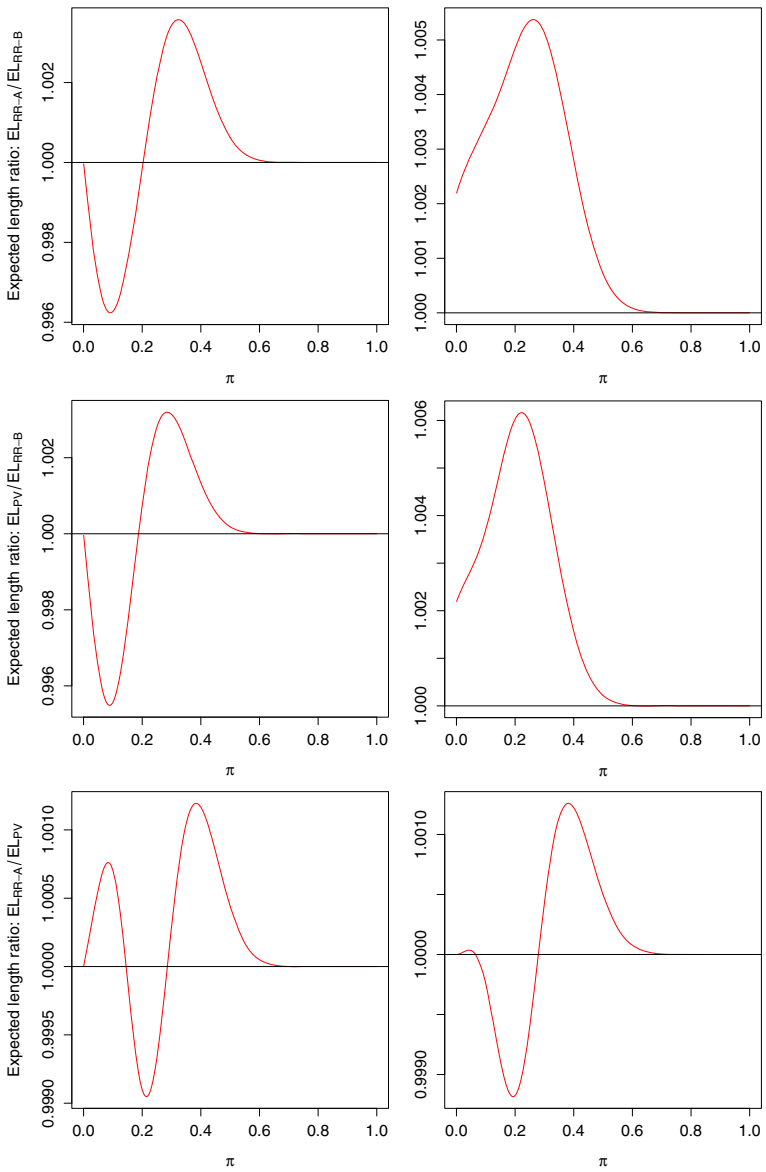

Fig. 5 Expected length comparison among the RR-A approach, the RR-B approach, and the PV approach, for the adaptive two-stage design with design parameters $(\pi_0, \pi_1, \alpha, \beta) = (20\%, 40\%, 0.05, 0.1)$ when the AdaptiveS design is used. The sample space G_2 (*left side*) and the subsample space $G_2(CI)$ (*right side*) are used to compute the expected lengths

RR-Score approach have very similar lower limits, and they are smaller than others. The RR approach has the largest lower limit among these approaches.

We also use this example to explain the reason why the RR approach is not exact. In the RR approach, the rejection region includes the extreme outcomes whose first stage and second stage responses are at least as large as the observed data. If the observed data is $X_1 = 11$ and $X_2 = 13$, then the sample point $(X_1 = 12, X_2 = 11)$ does not belong to the observed data's rejection region since $(12 + 11)/(22 + 35) < (11 + 13)/(22 + 35)$. By using the RR approach, the calculated 95% lower limit for $(X_1 = 12, X_2 = 11)$ is 0.362, which is larger than that of the observed data's. It follows that

$$(X_1 = 12, X_2 = 11) \notin \{(X_1, X_2) : L(X_1, X_2) < \pi | \pi = 0.343\}.$$

Thus, the sample point $(X_1 = 12, X_2 = 11)$ does not belong to the observed data's confidence set, neither. Therefore, the coverage probability at $\pi = 0.343$ could be less than the nominal level. In other words, in a two-stage design, inverting the p-value function is exact only when the sample space can be ordered completely by a test statistic or a measurement.

Discussion

In addition to the proposed confidence interval for adaptive two-stage designs to make statistical inference, unbiased response rate estimate and exact p-value calculation are two important future research topics. Due to the multi-stage nature of the design, the naive estimate that divides the total number of responses by the total number of participants, is biased. Jung and Kim [21] proposed an unbiased response rate estimate for traditional non-adaptive two-stage designs where the second stage sample size is a constant for a trial proceeding to the second stage. In adaptive two-stage design settings, the second stage sample size is a non-increasing function of the responses from the first stage. We consider this as future work to further develop an unbiased response estimate and exact p-value calculations in adaptive two-stage design settings.

Conclusions

Multiple adaptive two-stage designs have been proposed for use in practice, but these is limited research on analyzing the data once an adaptive study is finished. This article proposes three approaches to construct one-sided lower limits. Among these three new approaches, two approaches guarantee coverage probability. We compare these two exact intervals with another three existing approaches with regards to the two commonly used criteria: simple average length and expected length. The RR-A approach and the PV approach have similar performance with regards to these two criteria, although the RR-A approach performs slightly better than the PV approach under the AL criterion. The RR-B approach has a slightly longer average length than the other two approaches when all sample points are used in the calculation, and their difference becomes negligible when the subsample space $G_2(CI)$ is used. In addition, the RR-B approach generally has a shorter expected length than the RR-A approach and the PV approach by using the sample space $G_2(CI)$ [22–25].

Acknowledgements
We would like to thank the comments from two reviewers and associate editor, who help us to improve the manuscript.

Funding
Shan's research is partially supported by grants from the National Institute of General Medical Sciences from the National Institutes of Health: P20GM109025, P20GM103440, and 5U54GM104944. Zhang's work was supported by the Zhejiang Provincial Natural Science Foundation of China (grant no. LY15F020001) and the National Natural Science Foundation of China (grant no. 61170099).

Availability of data and materials
This is a manuscript to develop novel study designs, therefore, no real data is involved.

Authors' contributions
The idea for the paper was originally developed by GS and TJ. GS and HZ computed exact confidence intervals for adaptive one-arm two-stage designs in this paper. GS, HZ and TJ drafted the manuscript, revised the paper critically and approved the final version.

Authors' information
Guogen Shan: Epidemiology and Biostatistics Program, Department of Environmental and Occupational Health, School of Community Health Sciences, University of Nevada Las Vegas, 89154 Las Vegas, NV, USA. Hua Zhang: School of Computer and Information Engineering, Zhejiang Gongshang University, Hangzhou, 310018 Zhejiang, China. Tao Jiang (Corresponding author): Department of Statistics, Zhejiang Gongshang University, Hangzhou, 310018 Zhejiang, China.

Competing interests
The authors declare that they have no competing interests.

Consent for publication
Not applicable.

Ethics approval and consent to participate
Not applicable.

Author details
[1]Epidemiology and Biostatistics Program, Department of Environmental and Occupational Health, School of Community Health Sciences, University of Nevada Las Vegas, Las Vegas, NV 89154, USA. [2]School of Computer and Information Engineering, Zhejiang Gongshang University, Hangzhou, Zhejiang 310018, China. [3]Department of Statistics, Zhejiang Gongshang University, Hangzhou, 310018 Zhejiang, China.

Received: 6 October 2016 Accepted: 23 January 2017
Published online: 06 February 2017

References

1. Simon R. Optimal two-stage designs for phase II clinical trials. Control Clin Trials. 1989;10(1):1–10.
2. Berry DA. Adaptive clinical trials: the promise and the caution. J Clin Oncol. 2011;29(6):606–9. doi:10.1200/jco.2010.32.2685.
3. Yin G. Clinical trial design: Bayesian and frequentist adaptive methods, 1st edn: Wiley; 2012. http://www.worldcat.org/isbn/0470581719.
4. Lin Y, Shih WJ. Adaptive two-stage designs for single-arm phase IIA cancer clinical trials. Biometrics. 2004;60(2):482–90.
5. Shan G. Exact statistical inference for categorical data: Academic Press; 2016. http://store.elsevier.com/Exact-Statistical-Inference-for-Categorical-Data/Guogen-Shan/isbn-9780081006818/.
6. Shan G, Wilding GE, Hutson AD, Gerstenberger S. Optimal adaptive two-stage designs for early phase II clinical trials. Statist Med. 2016;35(8):1257–66. doi:10.1002/sim.6794.
7. Zhao J, Yu M, Feng X-PP. Statistical inference for extended or shortened phase II studies based on Simon's two-stage designs. BMC Med Res Methodol. 2015;15(1):48. doi:10.1186/s12874-015-0039-5.
8. Koyama T, Chen H. Proper inference from Simon's two-stage designs. Stat Med. 2008;27(16):3145–54.
9. Jovic G, Whitehead J. An exact method for analysis following a two-stage phase II cancer clinical trial. Stat Med. 2010;29(30):3118–25.
10. Shan G. Exact confidence limits for the response rate in two-stage designs with over or under enrollment in the second stage. Stat Methods Med Res. 2016. In press.
11. Clopper CJ, Pearson ES. The use of confidence or fiducial limits illustrated in the case of the binomial. Biometrika. 1934;26(4):404–13. doi:10.1093/biomet/26.4.404.
12. Duffy DE, Santner TJ. Confidence intervals for a binomial parameter based on multistage tests. Biometrics. 1987;43(1):81–93.
13. Atkinson EN, Brown BW. Confidence limits for probability of response in multistage phase II clinical trials. Biometrics. 1985;41(3):741–4.
14. Rosner GL, Tsiatis AA. Exact confidence intervals following a group sequential trial: a comparison of methods. Biometrika. 1988;75(4):723–9. doi:10.1093/biomet/75.4.723.
15. Jennison C, Turnbull BW. Group Sequential Methods (Chapman & Hall/CRC Interdisciplinary Statistics), 1st edn: Chapman and Hall/CRC; 1999. http://www.worldcat.org/isbn/0849303168.
16. Shan G, Ma C. Unconditional tests for comparing two ordered multinomials. Stat Methods Med Res. 2016;25(1):241–54. doi:10.1177/0962280212450957.

17. Shan G, Ma C, Hutson AD, Wilding GE. An efficient and exact approach for detecting trends with binary endpoints. Stat Med. 2012;31(2):155–64. doi:10.1002/sim.4411.
18. Shan G, Ma C, Hutson AD, Wilding GE. Randomized two-stage phase II clinical trial designs based on Barnard's exact test. J Biopharmaceutical Stat. 2013;23(5):1081–90. doi:10.1080/10543406.2013.813525.
19. Scherer R. PropCIs: Various Confidence Interval Methods for Proportions. R package version 19–0. 2013.
20. Banerjee A, Tsiatis AA. Adaptive two-stage designs in phase II clinical trials. Stat Med. 2006;25(19):3382–95.
21. Jung S-HH, Kim KMM. On the estimation of the binomial probability in multistage clinical trials. Stat Med. 2004;23(6):881–96. doi:10.1002/sim.1653.
22. Wang W, Shan G. Exact confidence intervals for the relative risk and the odds ratio. Biometrics. 2015;71(4):985–995.
23. Shan G. Exact statistical inference for categorical data, 1st edn: Academic Press; 2015. http://www.worldcat.org/isbn/0081006810.
24. Shan G, Wang W. ExactCIdiff: an R package for computing exact confidence intervals for the difference of two proportions. R Journal. 2013;5(2):62–71.
25. Wang W. Exact optimal confidence intervals for hypergeometric parameters. J Am Stat Assoc. 2015;(512). In press. doi:10.1080/01621459.2014.966191.

Submit your next manuscript to BioMed Central and we will help you at every step:

- We accept pre-submission inquiries
- Our selector tool helps you to find the most relevant journal
- We provide round the clock customer support
- Convenient online submission
- Thorough peer review
- Inclusion in PubMed and all major indexing services
- Maximum visibility for your research

Submit your manuscript at
www.biomedcentral.com/submit